17 Extraction of Attribute Values

17.1 Test Corpora

For evaluating our approach, we have used two popular information extraction corpora, the *CMU Seminar Announcements Corpus* and the *Corporate Acquisitions Corpus*.[1] The two corpora have been chosen so as to cover two widely different areas from the range of texts that our approach should be able to handle (cf. Sec. 7.1.2):

The Seminar corpus contains 485 seminar announcements (plain text files) collected from university newsgroup; the contained texts can be considered as *semi-structured* (cf. Sec. 6.2), since they are generally informal, quickly written e-mail-style messages which generally start with a loosely structured header part (cf. Fig. 9.1 on p. 63 for a typical example).

The Acquisitions corpus, on the other hand, contains 600 articles about mergers and acquisitions from the *Reuters-21578* corpus. These newspaper articles are classical *free texts*: they are formally written, strictly grammatical and contain almost no structured information. Together, the two corpora hence cover a broad range of the challenges that an IE system may encountered; the fact that they are two of the most frequently used IE corpora allows a comprehensive comparison with other IE systems.

Another popular corpus is the *Job Postings* collection of Mary E. Califf [Cal98b] which consists of 300 job offers posted to a Usenet newsgroup. This corpus represents a kind of semi-structured texts similar to the Seminar corpus, while being far less frequently used. Hence we preferred the Seminar corpus for this kind of text.

In the case of the Seminar corpus, the task is to extract up to four attributes from each document (if present): SPEAKER, LOCATION, START TIME (STIME) and END TIME (ETIME) of a talk. The answer keys for this corpus comprise 485 START TIMEs, 464 LOCATIONs, 409 SPEAKERs, and 228 END TIMEs.

The Acquisitions corpus defines nine attributes describing corporate mergers or acquisitions which should be extracted (the numbers of answer keys in the corpus are given in parentheses):

- the official names of the parties to an acquisition: ACQUIRED (593), PURCHASER (545), SELLER (235);
- the corresponding abbreviated names: ACQABR (437), PURCHABR (445), SELLERABR (182);
- the location of the acquired company: ACQLOC (178);
- the price paid: DLRAMT (259);
- information about the status of negotiations: STATUS (453).

[1] Both available from the *RISE Repository* [RISa].

For each corpus, we have used the typical evaluation setup. A training/test split of 50/50 is used for both corpora (50% of the documents are used for training and the rest of evaluation); results are averaged over five (Seminar) or ten (Acquisitions) random splits. For the Acquisitions corpus, we have used the ten random splits that are predefined by the corpus; the Seminar corpus does not specify any predefined splits, so we had to generate our own random splits.[2]

Both tasks are based on the assumption that each document describes only a single relevant relation, i.e., there is only one talk respectively one merger or acquisition per document whose details should be extracted (*text-as-tuple*, cf. Sec. 9.1). Some documents in the Seminar corpus contain additional pre-announcements of further talks, but these *should not* be extracted (extracting them will count as error).

Accordingly, "one answer per attribute" (or "match-best", cf. Sec. 15.2) is the typical evaluation mode for both corpora: at most one instance of each attribute is to be extracted from each document; if there are several answer keys for an attribute in a document, it is sufficient to find one of them. If our system finds multiple extraction candidates, it selects the most probably one. For the Acquisitions corpus, we will also give results for the "one answer per occurrence" or "match-all" evaluation mode to allow a comparison with the *ELIE* system which used that mode.

Unless stated otherwise, we will use the standard variant of our system, using *IOB2* as tagging strategy (cf. Sec. 10.2) and *Winnow* as classification algorithm (cf. Chap. 11). The metrics we will use have been introduced in Section 15.3; the reported average is always the *weighted average*. Except where noted otherwise, the reported performance figures are F-measure percentages—we will generally follow the usual convention of showing evaluation results as percentages, omitting the percent sign (96.5 is to be read as 96.5% or 0.965).

17.2 Evaluation Results for the Seminar Announcements Corpus

Table 17.1 shows the results reached by our system on the Seminar Announcements corpus. As discussed in Section 11.1.3, Winnow can be trained in two ways: either *incrementally*, which is faster and allows specific interactive annotation processes for reducing the human effort required to provide training data (an issue which we will investigate in the next chapter); or else via *batch* training, which is the conventional way of training information extraction systems and will generally lead to superior results since the classifier can make better use of the available training data. The table shows the results for both training regimens. Since a visual representation allows a more intuitive interpretation of results, the reached results with either of the training regimens are also shown graphically in Fig. 17.1.

[2] We did this once and then used the same set of splits for all subsequent tests. The lack of predefined splits is a weakness of the Seminar corpus since it means that differences in results reached by various systems might be partially caused by differences in the used splits. Malicious users could even improve the reported results of their system by repeatedly generating new random splits and reporting the results for the best set of splits.

Incremental Training			
	Precision	**Recall**	**F-measure**
etime	96.5	96.0	96.3
location	83.5	76.9	80.1
speaker	84.8	77.5	81.0
stime	99.3	99.3	99.3
Average	90.5	86.7	88.5
Batch Training			
	Precision	**Recall**	**F-measure**
etime	97.1	97.1	97.1
location	88.0	76.2	81.7
speaker	89.3	81.8	85.4
stime	99.3	99.3	99.3
Average	93.1	87.7	90.2

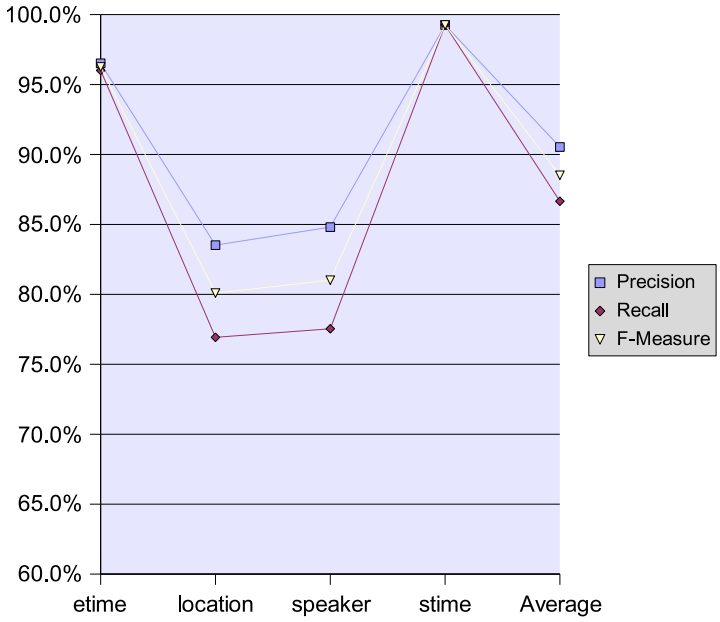
Table 17.1: Results on the Seminar Corpus

We can see that the START and END TIME attribute values are very easy for our system, which is not surprising since they are usually quite simple and regular. Identification of SPEAKERs and LOCATIONs is more difficult, but the system still reaches respectable F-measure values above 80%. In the case of these more difficult attributes, the system clearly favors precision over recall (for the START|END TIME attributes, both metrics are very near to each other)—we suppose that this bias towards precision is an effect of the classification-based nature of our system: since most tokens in a text are *not* part of any extraction, the classifier will always see far more negative training examples (not part of any attribute value) than positive ones, resulting in a tendency to choose the negative class instead of the positive one in case of dubious instances (where similar instances have been seen as representatives of both the positive class and the negative class).

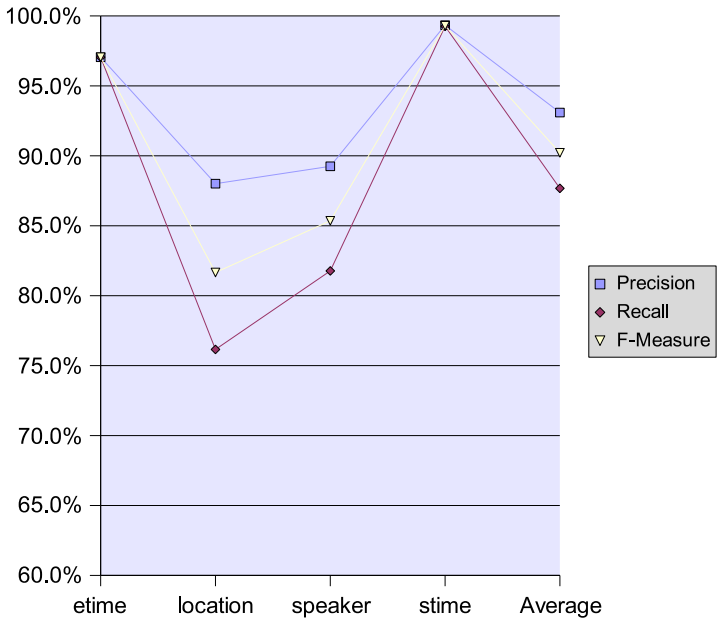
Switching from incremental to batch training improves the F-measure values for all attributes except STIME (where it is already at 99.3%, leaving very little room for improvement). Generally, it is more the precision (+2.6% on average) than the recall (+1.0% on average) which is improved by batch training—in case of LOCATION, the recall is actually reduced, but the larger gain in precision still results in a net F-measure improvement. The precision and recall improvements (or degradations) reached by switching from incremental to batch training are also shown graphically in Fig. 17.2.

Table 17.2 and Fig. 17.3 show a comparison of our system (referred to as TIE, "Trainable Information Extractor") with other approaches evaluated in the same way.[3] Our

[3] One other approach, BIEN [Pes03], is not directly comparable, since it uses an 80/20 split instead of 50/50. When run with an 80/20 split, the overall result of our system in incremental mode is 89.5%; BIEN reaches 88.9%.

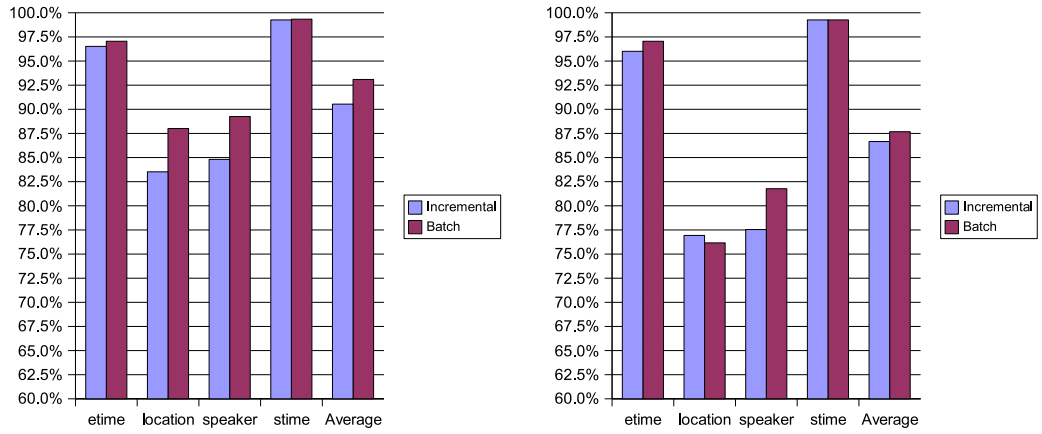


(a) Incremental Training



(b) Batch Training

Figure 17.1: Results on the Seminar Corpus



(a) Precision

(b) Recall

Figure 17.2: Seminar Corpus: Precision and Recall Improvements

Approach	**TIE**		**BWI**	**ELIE**		**(LP)^2**	**MaxEnt**	**MBL**	**SNoW-IE**	**CRF**
	Inc.	**Batch**		**L1**	**L2**					
Reference			[Fre00a]	[Fin04a]		[Cir01]	[Chi02]	[Zav03]	[Rot01]	[Sut05]
etime	96.3	**97.1**	93.9	87.0	96.4	95.5	94.2	96	96.3	96.0
location	80.1	81.7	76.7	84.8	**86.5**	75.0	82.6	**87**	75.2	85.3
speaker	81.0	85.4	67.7	84.9	**88.5**	77.6	72.6	71	73.8	76.3
stime	99.3	99.3	**99.6**	96.6	98.5	99.0	**99.6**	95	**99.6**	99.1
Average	88.5	90.2	83.9	88.8	**92.1**	86.0	86.9	86.6	85.3	88.7

Table 17.2: System Comparison on the Seminar Corpus (F-measure)

system competes very well with the other systems[4]—on average, already the results reached with incremental training (first column) are better than those of all other approaches (none of which supports incremental training), except for the *ELIE* system and a *CRF*-based approach.[5] ELIE is an approach which also uses token classification but in a different way (cf. Sec. 4.4.2) and which was developed independently at the same time as our own; Conditional Random Fields (CRFs) are a state-of-the-art statistical technique (cf. Sec. 4.3), the system shown here [Sut05] reaches its good results by integrating CRF models for named-entity recognition (cf. Sec. 14.1).

When using batch training instead of incremental training (second column), our system surpasses the CRF system and the first level of the ELIE system by more than 1%, while the results of the second level of ELIE remain better. As described in Sec. 4.4.2, ELIE uses token classification with Support Vector Machines in a two-level

[4] The table shows only results of the *best* other system evaluated on the Seminar corpus which we are aware of. There are many published results reached by other systems which are worse than those listed here; we have omitted them to keep the size of the table feasible.

[5] When judging from the published figures. It is not possible to determine whether performance differences to other systems are actually statistically significant since this would require detailed test results of the other systems which are not available.

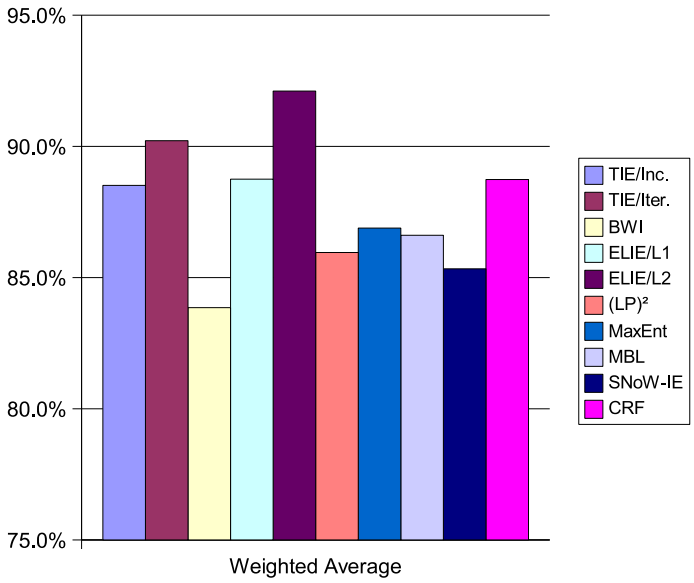


Figure 17.3: System Comparison: F-measure Averages on the Seminar Corpus

approach, while our system so far is limited to a single level. We might to able to reach similar improvements by using our system with the *BE* tagging strategy as they do (cf. Chap. 19) and adding a second level similar to theirs, but we did not try this since mimicking other people's ideas was not our goal.

The reported results are all from *trainable* systems—mainly statistical ones, while two (BWI and $(LP)^2$) use rule-learning. In the past, domain-specific rule-based systems haven often been able to outperform trainable approaches. However, for the evaluation corpora we are using we not aware of comparable or superior results reached by static, handcrafted systems.

17.3 Evaluation Results for the Corporate Acquisitions Corpus

Table 17.3 show the results of our system for the Corporate Acquisitions Corpus. Compared to the Seminar corpus, the reached results are very poor. The average F-measure is only 48% with incremental and 52% with batch training—probably unacceptable for any serious application.

Figure 17.4 graphically shows the results for both training regimens—again we see the tendency of our system to favor precision over recall, already for incremental and even stronger for batch training. The F-measure improvements of batch training compared to incremental training are large: typically about 3%–5%, more than 10% for the DLRAMT (dollar amount). The precision and recall differences between the two training modes are shown in Fig. 17.5. Similar to the Seminar corpus, it is especially the precision which is improved, the several cases by 10%–20% (ACQABR, ACQLOC, SELLERABR); while the recall actually drops in various cases (ACQABR, ACQ-

Incremental Training			
	Precision	**Recall**	**F-Measure**
acqabr	54.9	48.9	51.7
acqloc	50.3	18.7	27.3
acquired	62.7	40.5	49.2
dlramt	68.7	54.7	60.9
purchabr	54.7	55.9	55.3
purchaser	60.7	45.0	51.6
seller	54.4	17.1	26.0
sellerabr	40.4	17.1	24.0
status	58.2	48.7	53.0
Average	57.7	42.5	48.0
Batch Training			
	Precision	**Recall**	**F-Measure**
acqabr	65.7	47.3	55.0
acqloc	64.7	17.4	27.4
acquired	66.9	44.6	53.5
dlramt	76.7	67.3	71.7
purchabr	62.7	54.2	58.1
purchaser	66.8	47.7	55.7
seller	62.3	21.3	31.8
sellerabr	61.5	16.3	25.8
status	63.2	51.7	56.9
Average	65.7	44.8	52.1

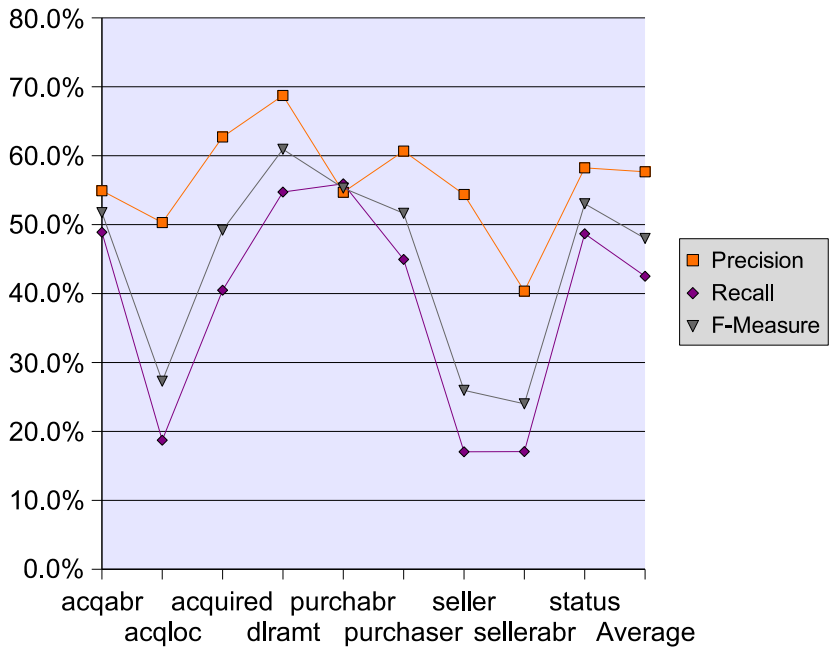
Table 17.3: Results on the Acquisitions Corpus

LOC, PURCHABR, SELLERABR) and is improved less strongly than the precision in the other cases.

An exception is the DLRAMT attribute, for which the recall (+12.6%) improves even more than the precision (+8.0%), resulting in the especially strong F-measure improvement noted above. In case of the ACQLOC attribute, a comparative small drop in recall (−1.3%) is sufficient to cancel out a large increase in precision (+14.4%), indicating how unfavorable the F-measure as the *harmonic* mean (cf. Sec. 15.3) judges such extremely unbalanced values.

Table 17.4 and Fig. 17.6 show the results of our system (TIE) compared to other approaches evaluated on the same corpus. Since the *ELIE* system has been tested in "match-all" instead of "match-best" mode (cf. Sec. 15.2), we list results in both modes. In both modes, TIE is clearly better than the other approaches (including ELIE), even when used with incremental training.

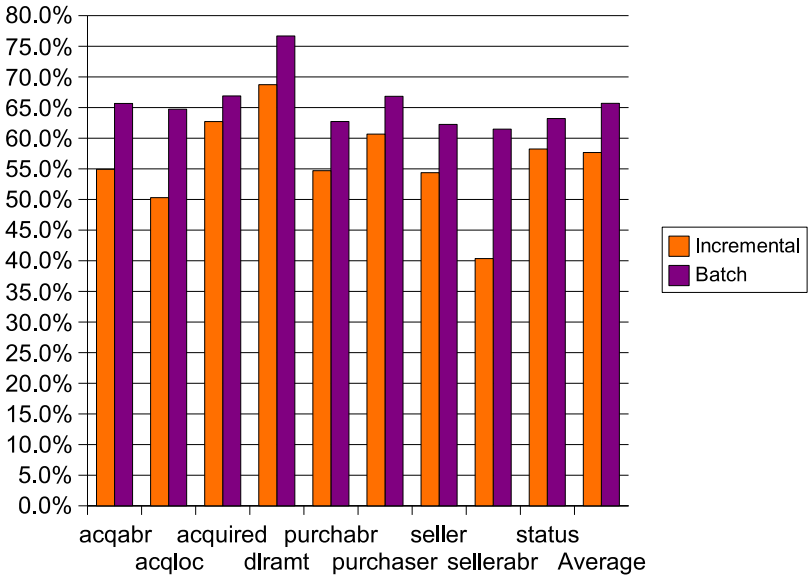
Results for the Acquisitions corpus are far worse than those for the Seminar corpus. This does not just hold for our system, but also for the other systems evaluated on both corpora—apparently the Acquisitions task is generally more "difficult" than the



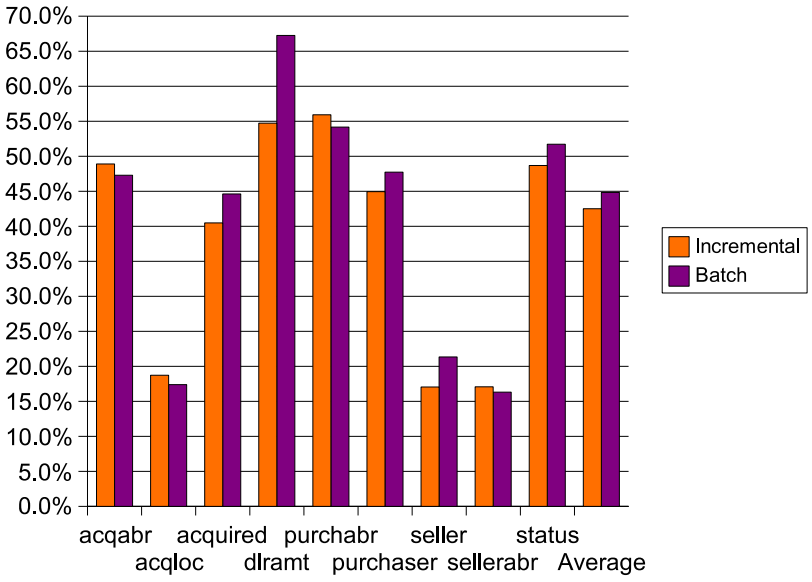
(a) Incremental Training

(b) Batch Training

Figure 17.4: Results on the Acquisitions Corpus



(a) Precision

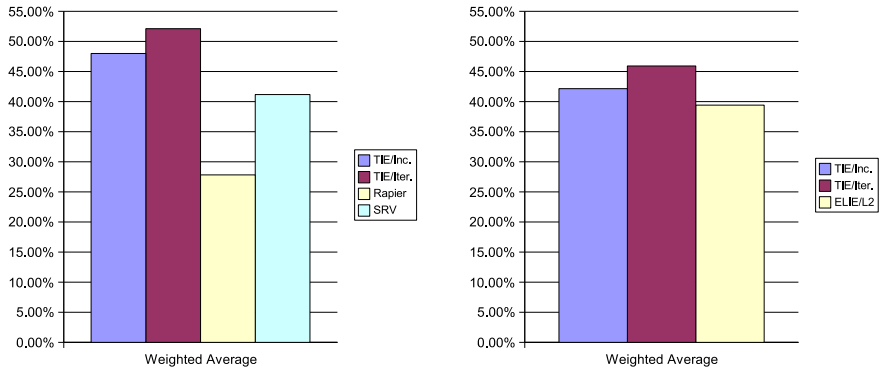


(b) Recall

Figure 17.5: Acquisitions Corpus: Precision and Recall Improvements

Mode	**Match-best**				**Match-all**		
Approach	**TIE**		**Rapier**	**SRV**	**TIE**		**ELIE/L2**
	Inc.	**Batch**			**Inc.**	**Batch**	
Reference			[Cal98b]	[Fre98a]			[Fin04b]
acqabr	51.7	**55.0**	26.0	38.1	42.7	**43.7**	39.7
acqloc	27.3	**27.4**	24.2	22.3	23.4	23.9	**34.4**
acquired	49.2	**53.5**	28.8	38.5	44.7	**49.2**	43.5
dlramt	60.9	**71.7**	39.3	61.8	59.4	**70.8**	59.0
purchabr	55.3	**58.1**	24.0	48.5	38.6	**40.5**	28.7
purchaser	51.6	**55.7**	27.7	45.1	48.4	**52.6**	46.2
seller	26.0	**31.8**	15.3	23.4	23.6	**28.7**	15.6
sellerabr	24.0	**25.8**	8.6	25.1	14.5	**16.4**	13.4
status	53.0	**56.9**	41.3	47.0	52.5	**56.3**	49.7
Average	48.0	**52.1**	27.8	41.2	42.1	**45.9**	39.4

Table 17.4: System Comparison on the Acquisitions Corpus (F-measure)



(a) Match-best Evaluation

(b) Match-all Evaluation

Figure 17.6: System Comparison: F-measure Averages on the Acquisitions Corpus

Seminar task. Without a deeper analysis (which we will partially perform in Chap. 21), we can identify three general factors that are likely to contribute to the bad results:

1. *Insufficient training data:* is is noticeable that the three attributes with the worst results are also the attributes with the lowest number of answer keys in the corpus: there are only 178/182/235 ACQLOC/SELLERABR/SELLER instances in the corpus, and the F-measure values of these attributes are lower by a striking difference of more than 20% than those of all other attributes in the corpus (for both training modes). The attribute with the fourth-lowest number of answer keys is DLRAMT (259 answer keys). This attribute specifies the dollar amount, i.e. the price paid for a acquisition, an attribute whose values are frequently numerical and usually short and regular, similar to the START and END TIME

attributes in the Seminar corpus. Still the F-measure results for this attribute, while respectable ($\approx 61\%$ for incremental, 72% for batch training) are very far away from the excellent results $> 96\%$ reached for the Seminar START|END TIME attributes, also indicating a lack of sufficient training data.

2. Differences in the *kinds of attributes* to extract: the values of the four attributes to extract from the Seminar corpus tend to be comparatively short, their meaning is clearly defined, and there is little risk of confusion between values of different types (except maybe for the START and END TIME attributes, but this risk is reduced by the fact that they tend to always appear in the same order, if both are present). The Acquisitions corpus, on the other hand, defines *six* different types of company names (the full and the abbreviated names of the three kinds of companies that can be involved), and the system is expected to be able to correctly differentiate between all of them—we can expect this to be a serious hurdle. Also, especially the full names of companies tend to longer than the person names and locations to identify in the Seminar corpus.
   The STATUS attribute is only vaguely defined; is comprises sometimes a single word, sometimes a whole phrase that is meant to somehow describe the status of negotiations—the vagueness in both meaning and form probably makes it hard for algorithms to detect suitable patterns and to exactly predict the attribute value chosen by the the human annotator. The low number of answer keys for the two remaining attributes (ACQLOC and DLRAMT) was already discussed above as another likely cause of problems.
3. Differences in the *kinds of texts* in the corpora. It is possible that *free texts* such as those in Acquisitions corpus are *generally* more difficult than *semi-structured* texts such as those in the Seminar corpus, and/or that there are *specific* difficulties in the way the articles forming the Acquisitions corpus are written. For now these are mere conjectures, but we will find some more evidence especially for the second conjecture when analyzing the kinds of mistakes that occur (Chap. 21).

Though a general study on what makes tasks harder or easier for automatic extraction is outside the scope of this work, it is evident that this is an important question for future work.